BSC-HIRLAM collaboration: HARMONIE code profiling current status

Xavier Yepes-Arbós
Mario C. Acosta
Kim Serradell

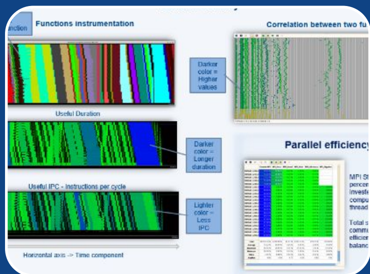
26/11/2019

HARMONIE System Working Week, MET Norway

Who we are

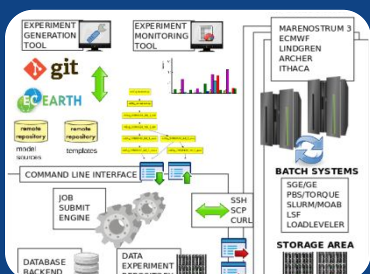
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Computational Earth Sciences Group



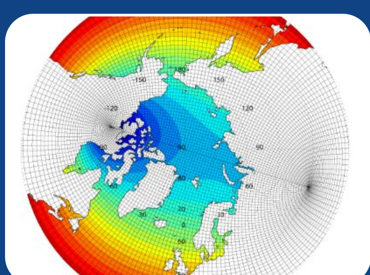
Performance Team

- Provide HPC Services (profiling, code audit, …)
- Apply new computational methods



Models and Workflows Team

- Development of HPC user-friendly software framework
- Support the development of atmospheric research software



Data and Diagnostics Team

- Big Data in Earth Sciences
- Provision of data services
- Visualization

Performance Team

- The necessary refactoring of numerical codes is given a lot of attention and is stirring a number of discussions
  - Computational performance analysis and new optimizations are needed for actual numerical models
  - Studying new algorithms for the new generation of high performance platforms (path to exascale)
- We are collaborating with several institutions on different projects working together in the same direction

is-enes
INFRASTRUCTURE FOR THE EUROPEAN NETWORK FOR EARTH SYSTEM MODELLING

immerse
IMPROVING OCEAN MODELS FOR THE COPERNICUS PROGRAMME

High Resolution Hirlam Limited Area Model

esiwace
CENTRE OF EXCELLENCE IN SIMULATION OF WEATHER AND CLIMATE IN EUROPE

MERCATOR OCEAN INTERNATIONAL

ETP 4 HPC EsDs

ESCAPE 2

PRACE

intel

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Roadmap

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

BSC-HIRLAM collaboration

- The BSC and the HIRLAM consortium signed a contract for a 1 year project to perform a complete code profiling of the HARMONIE-AROME model and the Data Assimilation system

- The project consists of two phases:
  - 1st: Basic profiling analysis
  - 2nd: Perform a complete analysis according to the results from the first phase

Scope of the phase 1

- Duration: 4 months
- Prepare selected configurations to be deployed with Extrae on cca/ccb at ECMWF, a Cray XC40 machine
- Perform a basic analysis of the HARMONIE-AROME Forecast model and the Data assimilation execution
  - Use different computational metrics: IPC, useful duration, MPI overhead, cache misses, etc
  - Identify the different parts of the trace with regard the code being executed
- Deliver a complete document with the results and feedback to decide the main goals for the profiling analysis of the phase 2

Scope of the phase 2

- Duration: 8 months after completion of phase 1
- Complete profiling analysis according to the results obtained from phase 1
- Training:
  - Prepare basic tutorials based on coarser HARMONIE configurations
  - Prepare a physical event to perform a training for the users
- Prepare complete documentation:
  - Online follow-up meetings to detect deviations and correct if needed
  - Write a final document describing all the tasks carried on
- Presence in the HIRLAM System group meetings: dissemination and feedback

Code profiling methodology

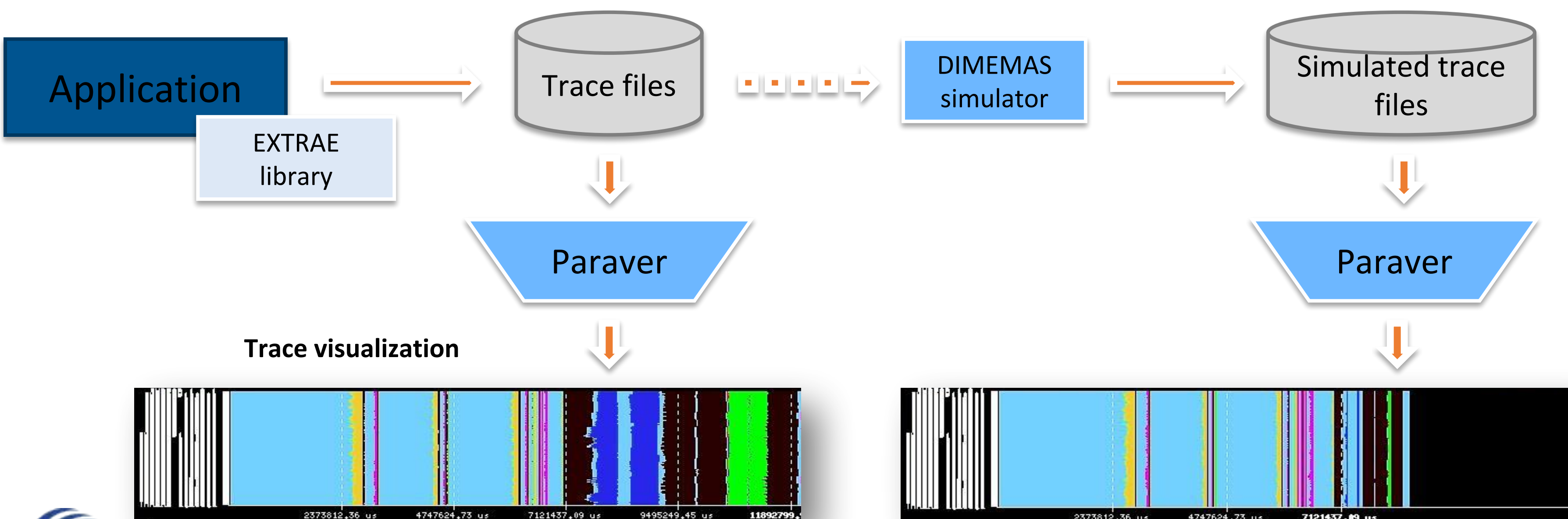
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Profiling methodology overview

- Scalability tests: MPI, OpenMP, Tiling
- Evaluate deployment efficiency: compilation flags
- Affinity tests: find a proper placement for MPI processes
- Profile analysis: user functions calls, statistics…
- Trace analysis: MPI, hardware counters, communication…
- Performance simulation: evaluate the code under machine changes
- Validation tests: check code correctness if optimized

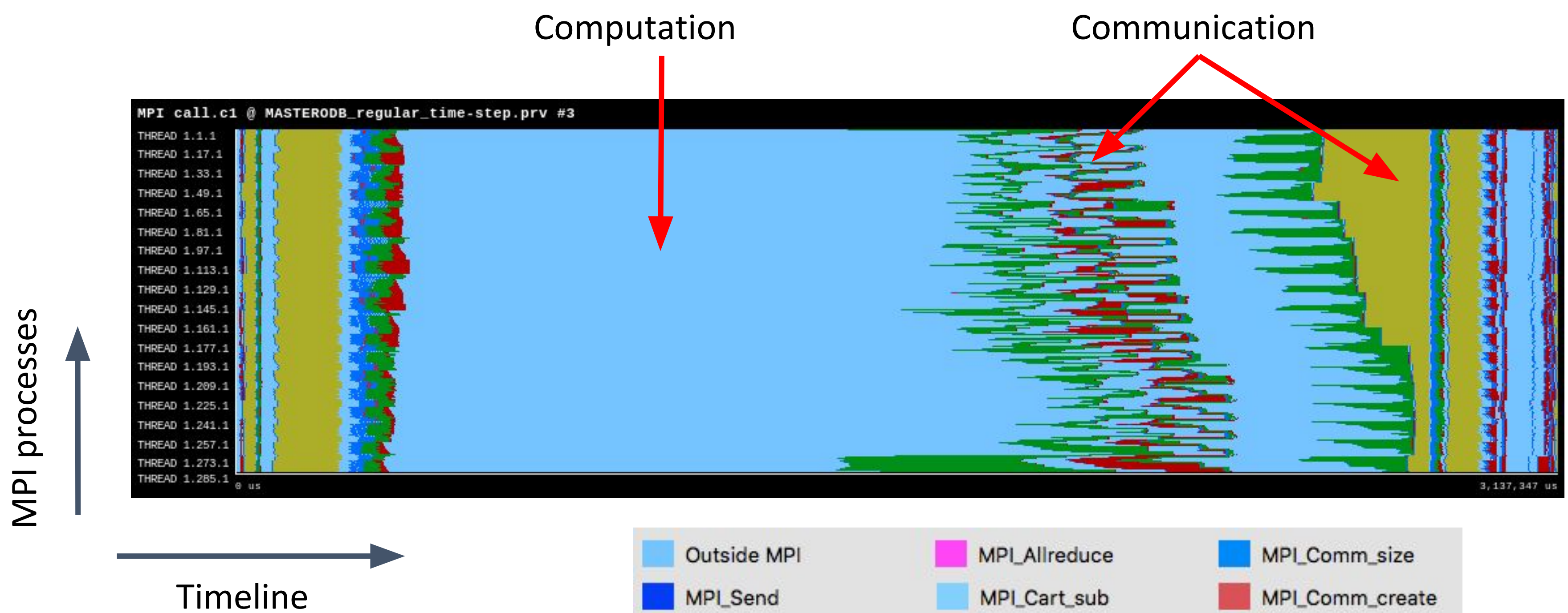
BSC performance tools

- Since 1991
- Based on traces
- Open Source: http://www.bsc.es/paraver
- **Extrae**: Package that generates Paraver trace-files for a post-mortem analysis
- **Paraver**: Trace visualization and analysis browser
  - Includes trace manipulation: Filter, cut traces
- **Dimemas**: Message passing simulator

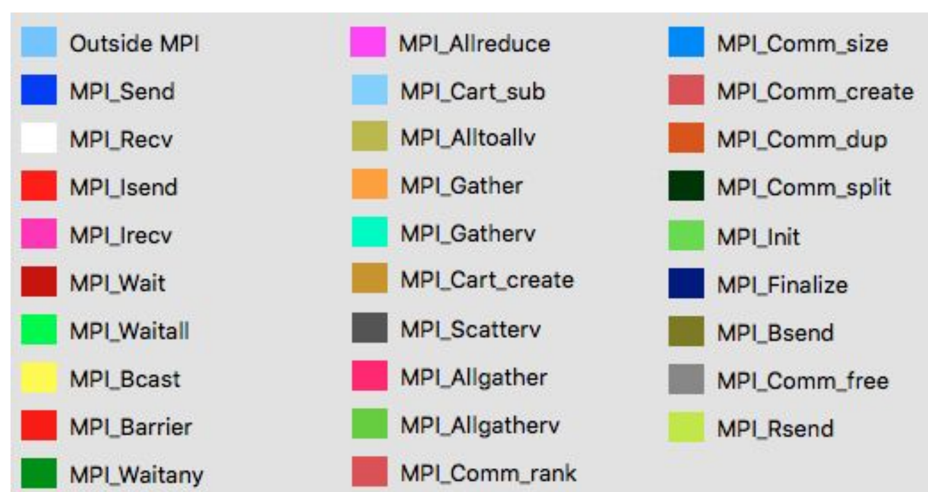


DIMEMAS generated trace. Target = ideal machine

How a trace looks like: basic overview



MPI call color legend



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Current status

Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación

Deployment

Common configuration for the different scenarios:

Branch	release-43h2.beta.5
Domain	METCOOP25C (2.5km and 65 vertical levels)
Configuration	default
Cluster	Cray XC40 (ECMWF cca)
Compiler	gcc/7.3.0

Deployment (2)

- Preliminary profiling analysis of this scenario:
  - 285 MPI & 1 OpenMP (default) & no I/O server

- And strong scalability tests for these ones:
  - 285 MPI & 1 OpenMP (default) & no I/O server
  - 143 MPI & 2 OpenMP correct job geometry & no I/O server
  - 72 MPI & 4 OpenMP correct job geometry & no I/O server
  - 285 MPI & 6 OpenMP correct job geometry & no I/O server

- No traces enabling the I/O server due to an Extrae issue

Extrae issue

- When we trace HARMONIE with the I/O server enabled, everything hangs

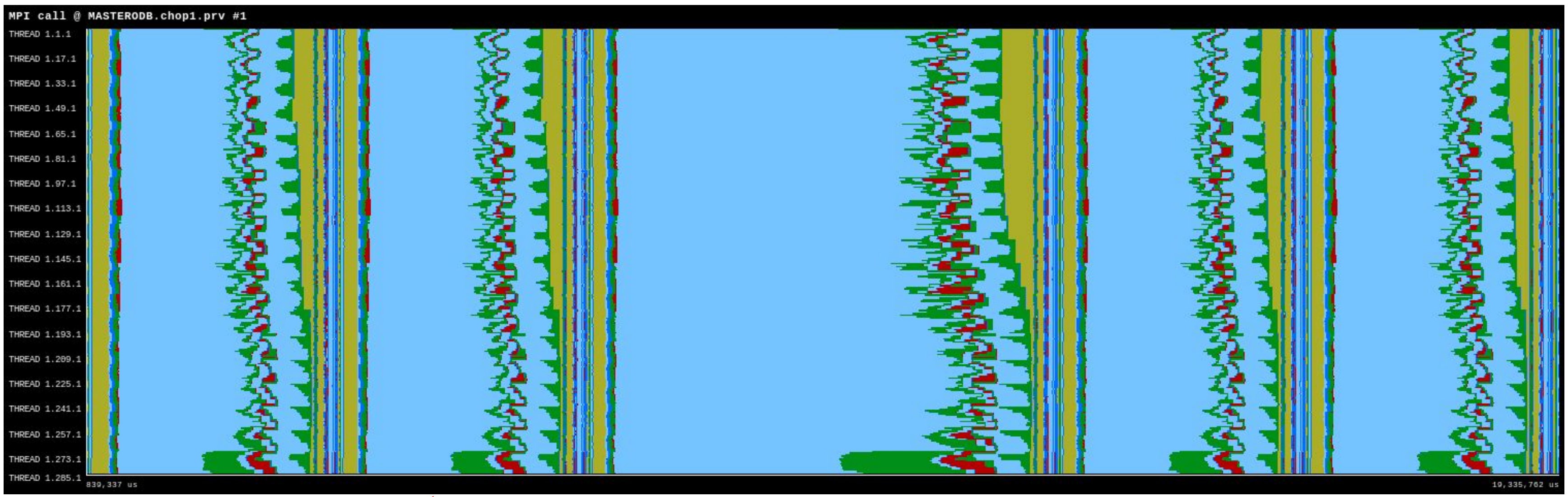
- There is a really rare issue on Extrae from an unknown source

- We are actively collaborating with the BSC tools developers to solve this issue as soon as possible

- However, this is not a problem now to profile a time step of HARMONIE

Structure of HARMONIE

- Cut trace containing 5 time steps from the profiled scenario
- Original trace is get from a 1-hour run -> about 30 GB

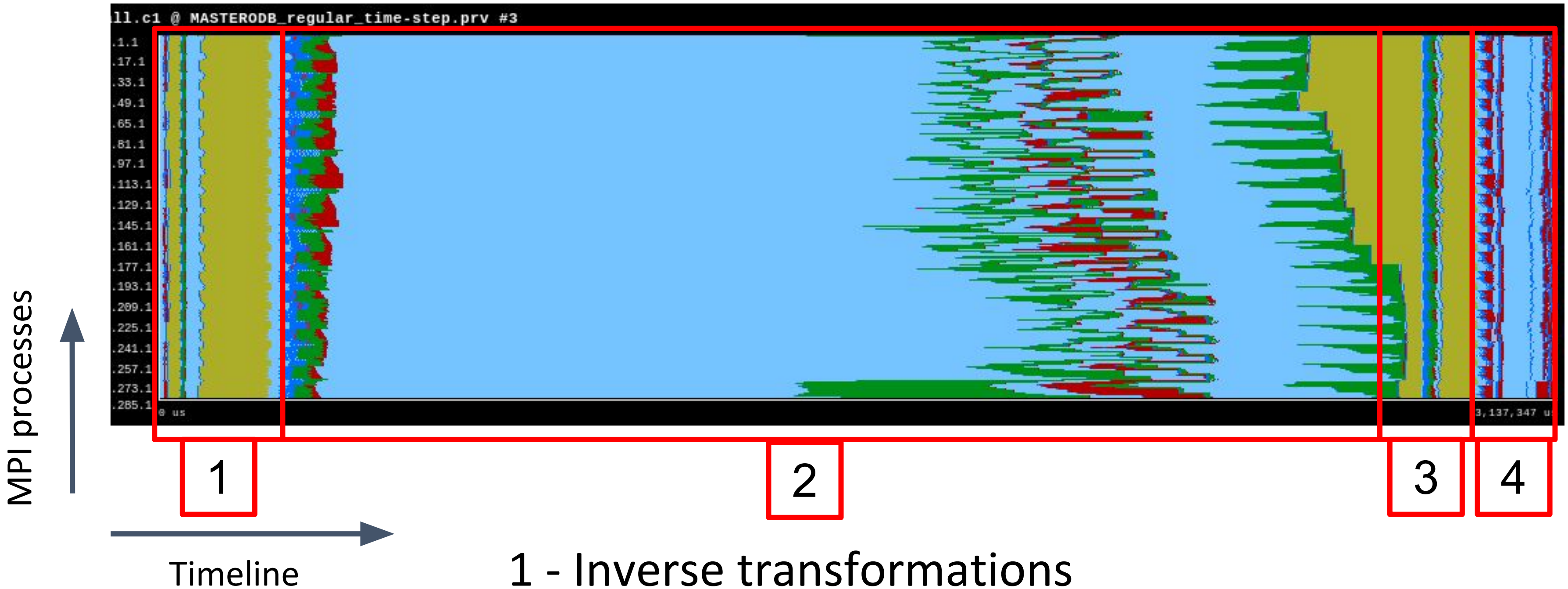


Regular time steps

Larger time step (it is probably computing radiation)

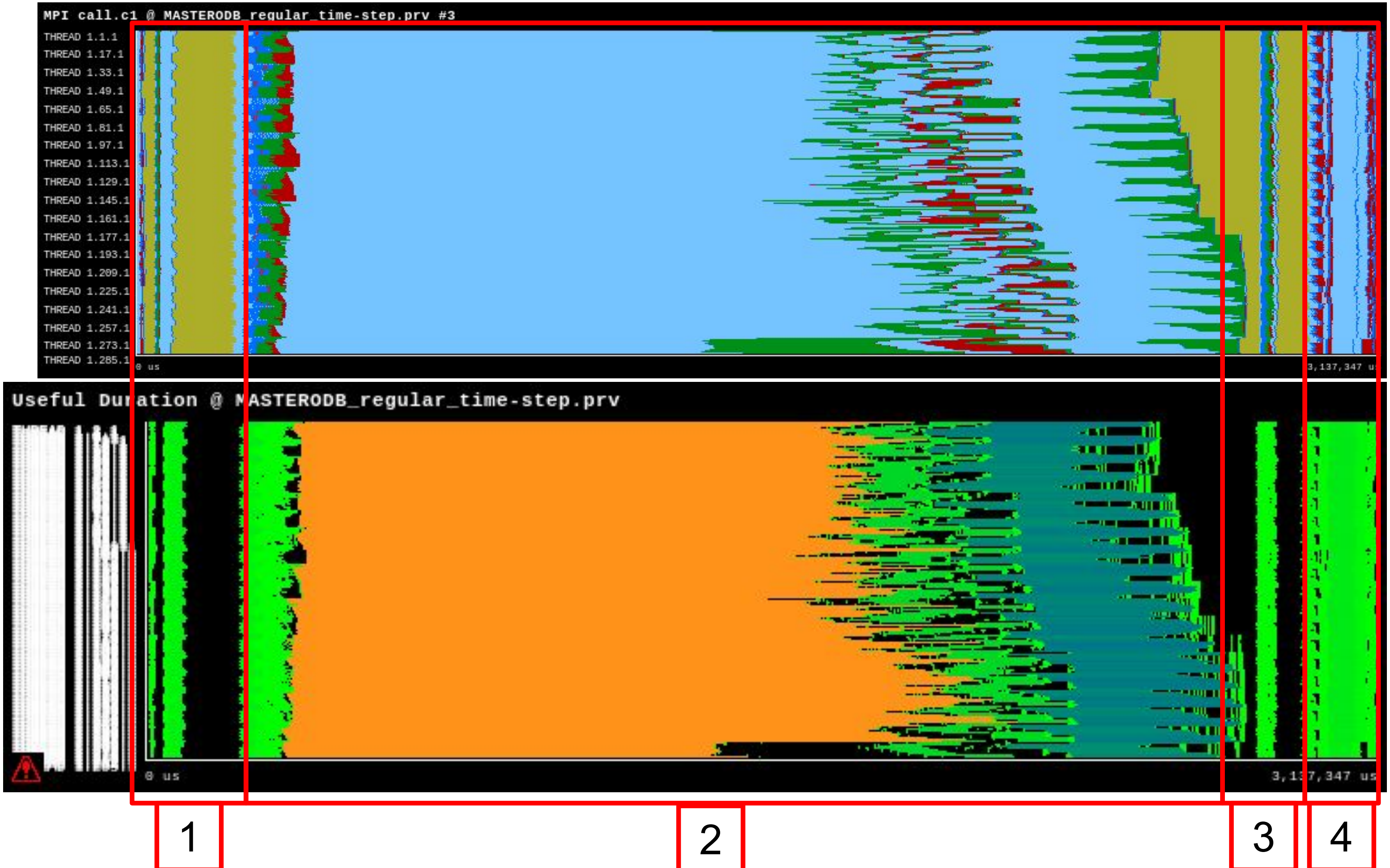
Structure of a regular time step

Duration: 3.14 seconds



1 - Inverse transformations
2 - Grid-point computations
3 - Direct transformations
4 - Spectral computations

Structure of a regular time step (2)



Strong scaling

nx*ny*th

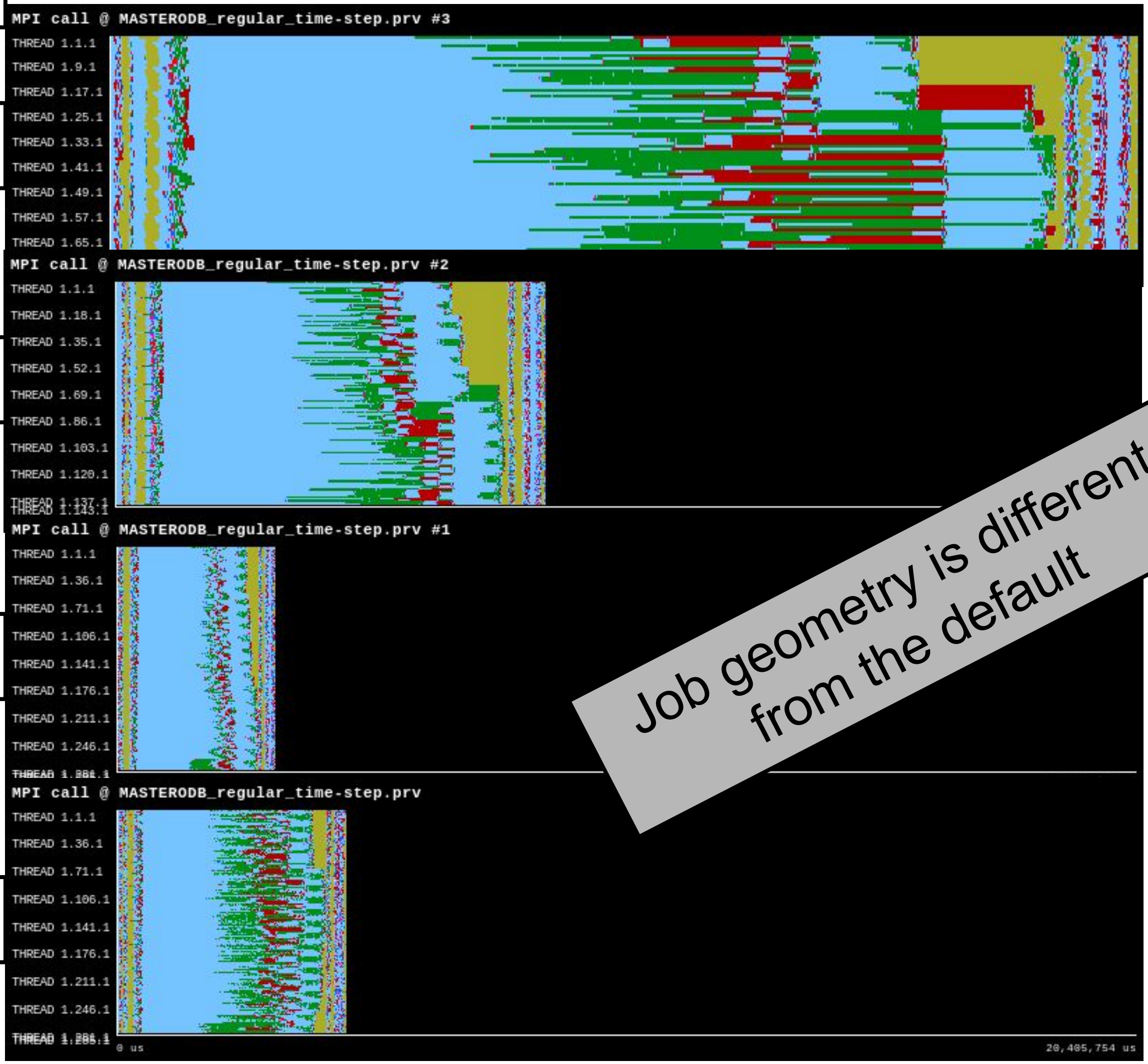
8*9*4 = 288

11*13*2 = 286

15*19*1 = 285

(default)

15*19*6 = 1710



Job geometry is different from the default

OpenMP scalability issues

- Preliminary scalability tests suggests that OpenMP is not scaling well

- It is necessary to investigate the reason

- One typical reason is due to the granularity of the OpenMP computational chunks:
  - If they are too fine -> overhead issues
  - If they are too coarse -> load imbalance issues

Job geometry

- Apparently, the job geometry is not properly set:
  - PBS clauses
  - aprun command

- The problem is that “EC_threads_per_task” is not properly set

- In addition, aprun command is supposedly set with “-cc cpu -ss”, but it isn’t

- MPI master process is run in an exclusive node, but using an I/O server, it shouldn’t be necessary. It will be tested in the future

Compilation flags

- On cca, GNU compiler uses -ffast-math. It is quite risky to use such aggressive floating point optimizations. Consider reproducibility tests

- If you usually run HARMONIE using only one OpenMP thread, consider removing the -fopenmp flag, because the OpenMP overhead can affect performance. It will be tested in the future

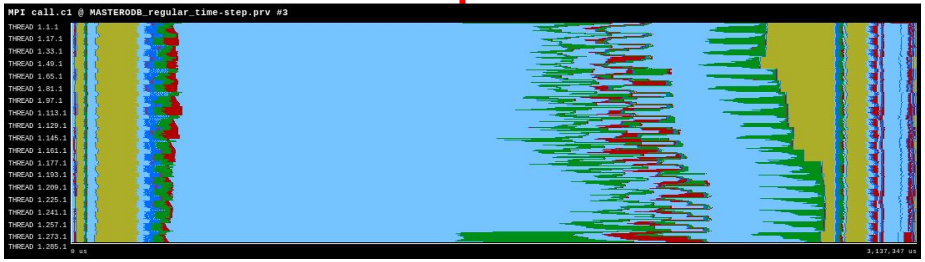
- We are currently investigating the automatic vectorization of the code, but it might be a good idea to use the -ftree-vectorize flag

Load balance of the time step

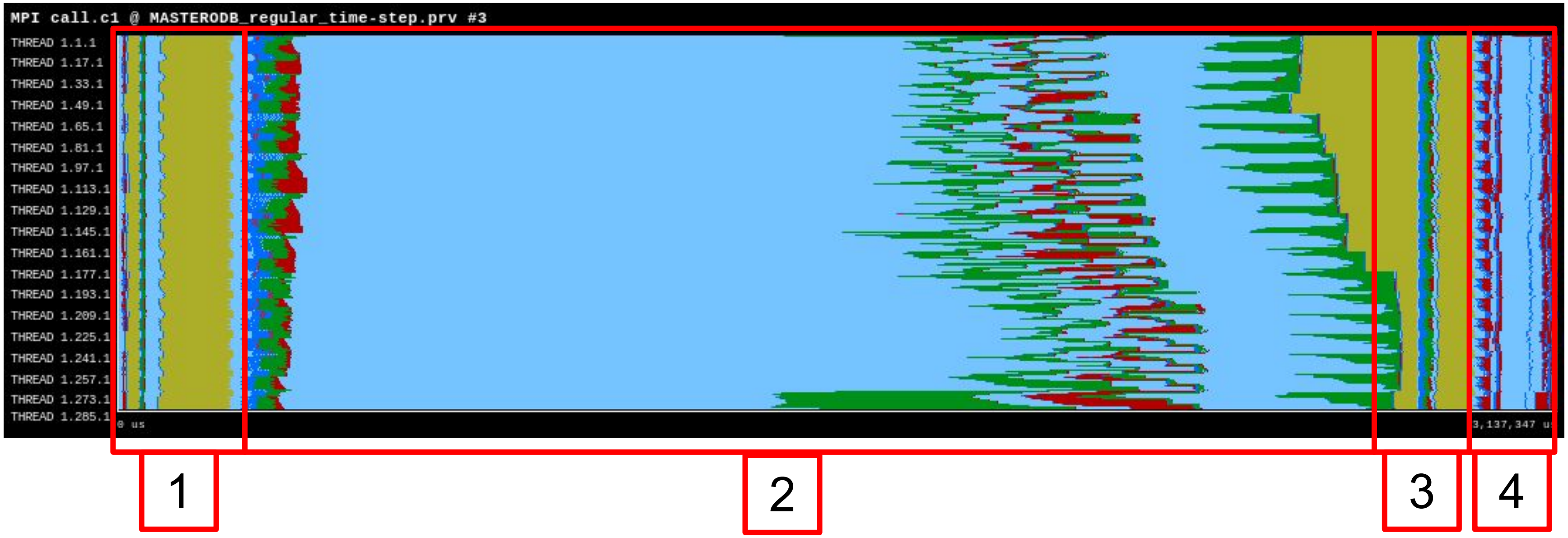
- MPI call profile of the time step:

	Outside MPI	MPI_Recv	MPI_Isend	MPI_Irecv	MPI_Wait	MPI_Alltoallv	MPI_Comm_size	MPI_Waitany
Total	21,001.26 %	4.27 %	88.61 %	29.68 %	1,149.04 %	3,636.14 %	58.84 %	2,532.16 %
Average	73.69 %	0.02 %	0.31 %	0.10 %	4.03 %	12.76 %	0.21 %	8.88 %
Maximum	86.12 %	0.12 %	0.96 %	0.20 %	11.13 %	17.85 %	0.25 %	24.84 %
Minimum	52.22 %	0.00 %	0.08 %	0.06 %	0.69 %	9.04 %	0.16 %	0.95 %
StDev	5.56 %	0.02 %	0.19 %	0.02 %	2.23 %	2.81 %	0.02 %	4.29 %
Avg/Max	0.86	0.13	0.32	0.52	0.36	0.71	0.82	0.36

- Parallel efficiency: 73.69%
- Communication efficiency: 86.12%
- Load balance: 86%



Load balance of the time step (2)



1 - Inverse transformations - good load balance (90%)

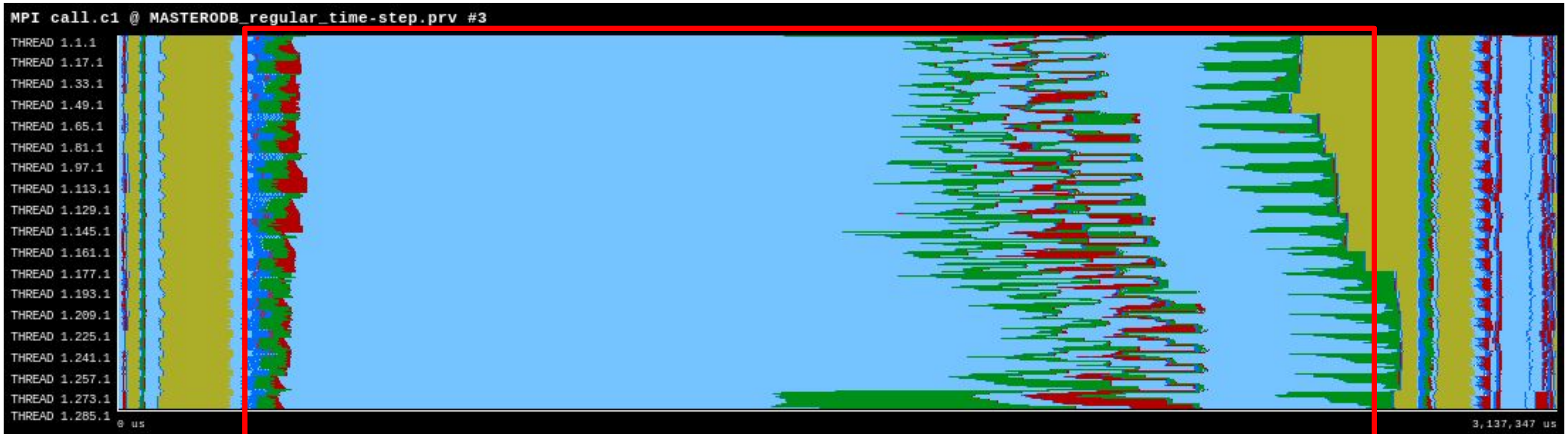
2 - Grid-point computations - some load imbalance (84%)

3 - Direct transformations - load balance is ok (87%)

4 - Spectral computations - load balance is ok (87%)

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Load imbalance of the grid-point part

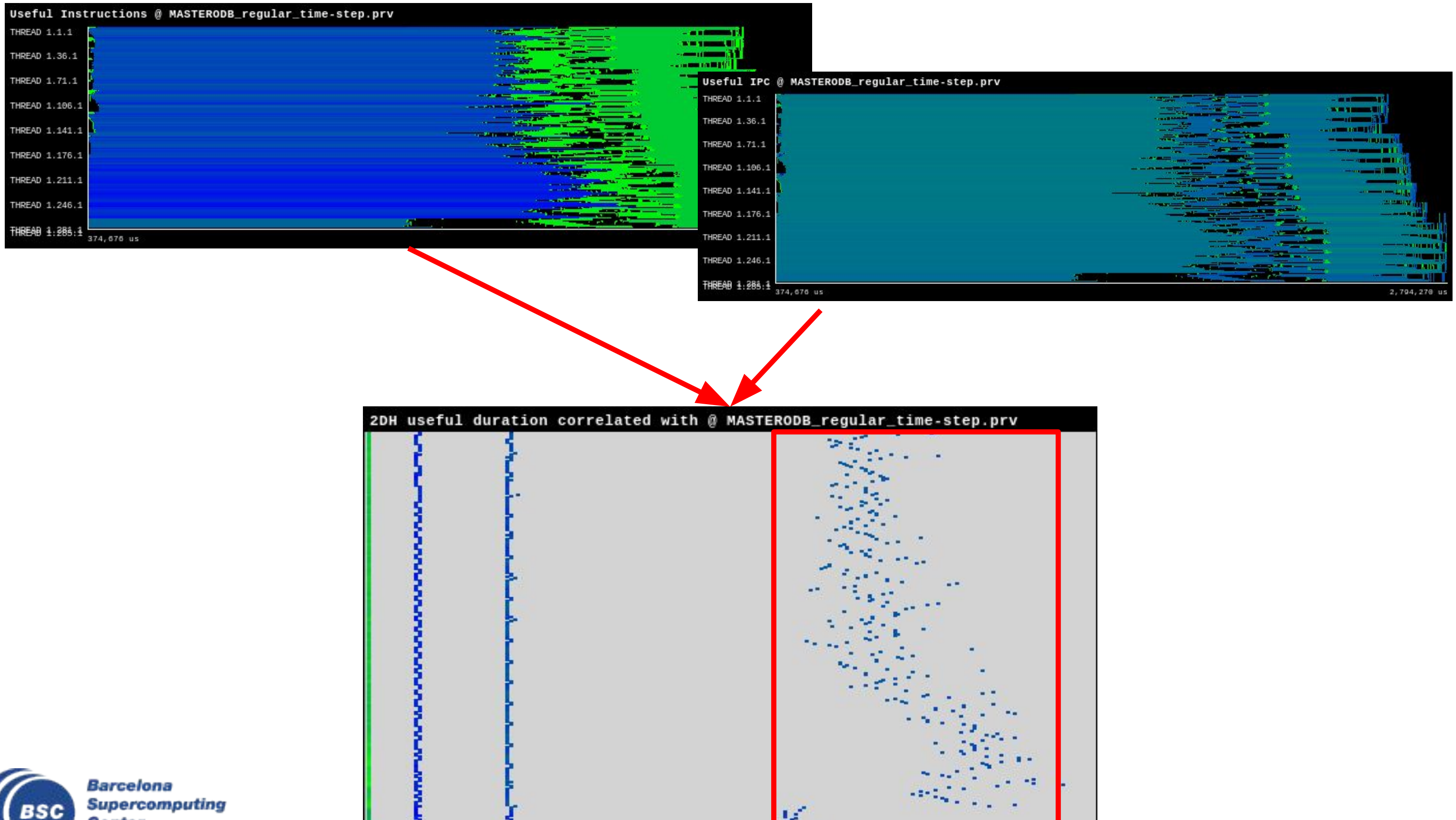


	Outside MPI	MPI_Isend	MPI_Irecv	MPI_Wait	MPI_Alltoallv	MPI_Comm_size	MPI_Waitany
Total	23,105.39 %	80.92 %	13.27 %	1,029.64 %	1,199.67 %	25.51 %	3,045.60 %
Average	81.07 %	0.28 %	0.05 %	3.61 %	4.54 %	0.09 %	10.69 %
Maximum	96.48 %	1.10 %	0.17 %	11.44 %	9.59 %	0.12 %	30.69 %
Minimum	59.04 %	0.03 %	0.02 %	0.12 %	0.00 %	0.05 %	0.80 %
StDev	6.61 %	0.22 %	0.02 %	2.61 %	3.32 %	0.01 %	5.37 %
Avg/Max	0.84	0.26	0.28	0.32	0.47	0.75	0.35

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

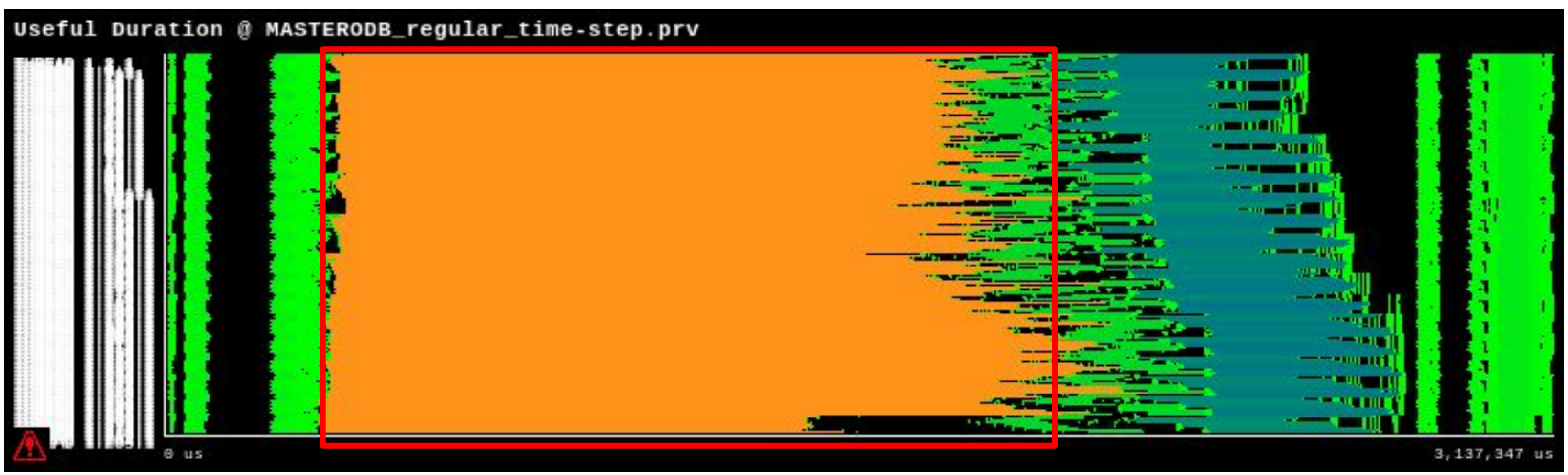
Source of the load imbalance

- Histogram of correlated instr. with IPC of the grid-point part
- Some processes have more workload (more inst. & same IPC)



Large computational phase

- Computation of the physics requires a lot of time
- Determine whether it is compute or memory bound:
  - Memory bound: high cache misses ratio and low IPC
  - Compute bound: low cache misses ratio and high IPC:
    - Low vectorization efficiency: not properly vectorized
    - High vectorization efficiency: optimal



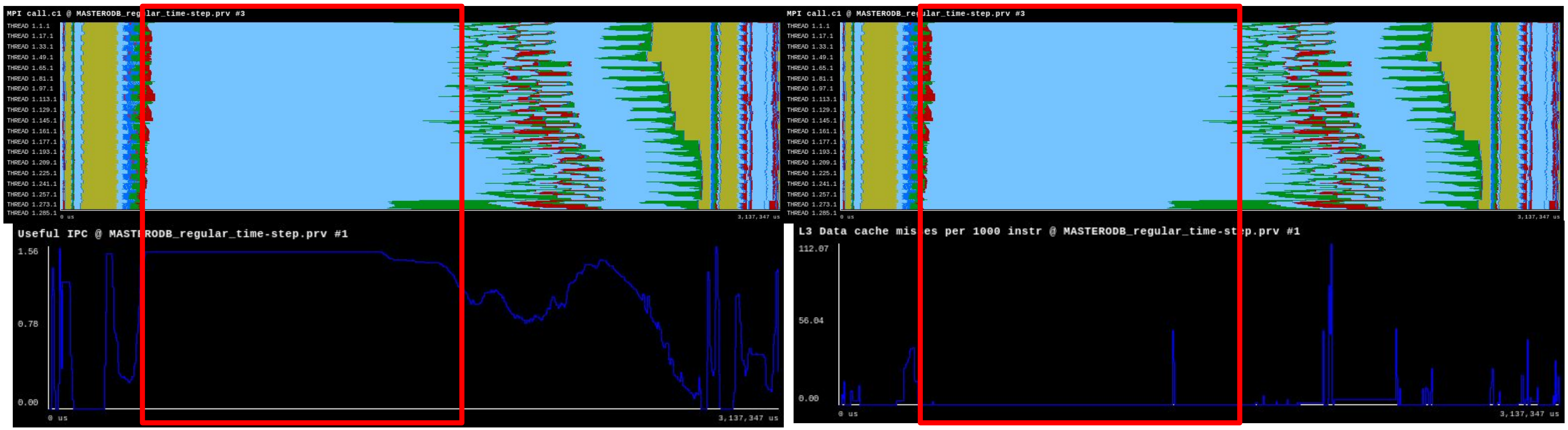
Shorter computation burst

Larger computation burst

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Large computational phase (2)

Aggregated IPC vs. aggregated L3 cache misses per 1000 instr.:



IPC > 1

L3 cache misses ratio < 1

Grid-point phase is
compute bound

Large computational phase (3)

- Since the grid-point phase is compute bound, we have to determine if vectorization is being properly applied or not

- However, we do not have this info on Paraver due to some missing PAPI counters. We will try to find a workaround as well as to analyze the GNU compiler report about vectorization (-fopt-info-vec-optimized)

What’s next?

Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación

Future work

- Perform a more complete MPI and OpenMP strong scalability tests (nx*ny*threads) and tracking study
- Compare 1 thread OpenMP vs. no OpenMP support
- Change job geometry (fix OpenMP setup, PBS clauses, explore binding (-cc), non-crossed memory allocation (-ss), avoid MPI master in an exclusive node)
- Change compilation flags (-O[2,3], -ftree-vectorize flag, -ffast-math, etc)
- Trace user functions, other PAPI counters, etc
- Dimemas simulations to evaluate the code under machine changes

Discussion

- Based on our experience, Intel achieves more performance than GNU. Is Intel compilation already working? If so, what to use?
  - Intel vs. GNU

- Daniel suggested that the load imbalance in the grid-point computation phase could be due to the type of HARMONIE grid. Could you provide more feedback to me?

- Default configuration is OK?

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

Thank you

xavier.yepes@bsc.es